Artificial Intelligence and Social Computing, Vol. 72, 2023, 89–99
https://doi.org/10.54941/ahfe1003277

Does Imageable Language Make Your Tweets More Persuasive?

Andy Bernhardt, Tomek Strzalkowski, Ning Sa, Ankita Bhaumik, and Gregorios Katsios

Rensselaer Polytechnic Institute, Troy, NY 12180, USA

ABSTRACT

Imageability is a psycholinguistic property of words that indicates how quickly and easily a word evokes a mental image or other sensory experience. Highly imageable words are easier to read and comprehend, and, as a result, their use in communications, such as social media, makes messages more memorable, and, potentially, more impactful and influential. In this paper, we explore the relationship between the imageability of messages in social media and their influence on the target audience. We focus on messages surrounding important public events and approximate the influence of a message by the number of retweets the message receives. First, we propose novel ways to determine an imageability score for a text, utilizing combinations of word-level imageability scores from the MRCPD+ lexicon, as well as word embeddings, image caption data, and word frequency data. Next, we compare these new imageability score functions to a variety of simple baseline functions in correlation between tweet imageability and number of retweets in the domain of the 2017 French Presidential Elections. We find that the imageability score of messages is correlated with the number of retweets in general, and also when normalized for topic and novelty; thus, imageable language is potentially more influential. We consider grouping tweets into imageability score ranges, and find that tweets within higher ranges of imageability scores receive more retweets on average compared to tweets within lower ranges. Lastly, we manually annotate a small number of tweets for imageability and show that our imageability score functions agree well with the human annotators when the agreement between human raters is high.

Keywords: Nlp, Imageability, Influence, Social media

INTRODUCTION

Imageability is a psycholinguistic property of words that indicates how quickly and easily a word evokes a mental image or other sensory experience. Seminal research has shown that the imageability of an individual word is related to how well it is encoded in memory (Paivio, 1969), and additionally involved an experimental scaling of 925 English nouns by students for concreteness, imagery, and meaningfulness measures, where imagery is defined "in terms of a word's capacity to arouse nonverbal images" (Paivio et al. 1968). Additionally, more recent research has shown that presence of highly imageable words is a strong indicator of metaphorical language in texts (Broadwell et al. 2013), as well as a variety of linguistic devices including metaphors,

similes, and metonyms. Such figurative language is considered more persuasive since it attempts to reduce complex and often abstract concepts to easily graspable, concrete ones, e.g., bureaucracy being compared to a maze, while also supplying an affective orientation, i.e., bureaucracy, like mazes, are hard to navigate and get out of. Highly imageable words are easier to read and comprehend, and, as a result, their use in communications, such as social media, makes messages more memorable, and, potentially, more impactful and influential.

Much of prior research in imageability and related psycholinguistic properties has been on the word or phrase level. Coltheart (1981) introduced the MRC psycholinguistic database (MRCPD), which includes 9,240 word-level imageability scores, among scores for other "syntactic, semantic, orthographic, and phonological properties", and this lexicon has been expanded by using WordNet synonyms and hyponyms and translated into a variety of languages (Liu et al. 2014; Liu et al. 2016). For example, the English MRCPD+ contains imageability scores for 116,151 English words, and has been evaluated structurally and also against imageability ratings from human subjects (Liu et al. 2014). Thanks to the availability of word-level imageability scores, many applications of imageability in longer texts, such as metaphor extraction, utilize these scores directly. In this paper, we extend word-level imageability to the sentence level, and propose novel ways to determine an imageability score for a text. Similarly to the word-level imageability, the imageability of a sentence can be defined by how easily the sentence evokes a mental image. This is related to the imageability of the individual words that make up the sentence, but also how these words relate within the context. As a result, challenges arise, such as the coherence of an image or images evoked by words in a sentence. For example, the words "cat" and "bicycle" are both relatively imageable, and the text "cat on a bicycle" can be visualized as well, but "cat bicycle" may be less evocative. We address these challenges in our development of imageability score functions for sentences.

Additionally, we are interested in exploring the relationship between the imageability of social media messages and their influence on the target audience. We focus on Twitter and the body of messages surrounding important public events, such as the 2017 French presidential elections, Covid-19 pandemic, and Russo-Ukrainian war. Such events are typically associated with multiple influence campaigns, where various actors, both official and clandestine, attempt to shape public opinion and potentially the outcomes of the event. It is not always easy to identify or measure the impact of an influence campaign, but one useful proxy is the response rate from the public, in terms of the number of retweets, or other direct reactions to campaign messages (Cha et al. 2010). In other words, messages that generate a large number of retweets can be considered more influential than those that receive few or none. However, predicting whether a particular message or trend will be popular, is challenging, since a large number of factors affect this, including the content/topic strength at the current time and poster's reach, as well as the novelty and believability of the message content (Bakshy et al. 2011). We address this when considering the relationship between imageability and

retweets in small subsets of tweets that are related to timeless topics as well as topics related to events.

In this paper, we propose novel ways to determine an imageability score for a text, utilizing combinations of word-level imageability scores from the MRCPD+ lexicon, as well as word embeddings, image caption data, and word frequency data. We compare these imageability score functions to a variety of simple baseline functions in correlation between tweet imageability and number of retweets in the domain of the 2017 French Presidential Elections, and consider case studies involving tweets related to specific topics or events. Lastly, we manually annotate a small number of tweets for imageability to spot-check the quality of one of the imageability score functions as an estimate of sentence-level imageability.

RELATED WORK

Recent research in imageability has leaned toward estimating the imageability in longer texts, such as sentences or image captions, and additionally addresses the idea of coherence of images evoked by words in a sentence.

Madden-Lombardi et al. (2015) experimentally provide human raters with pairs of sentences that are "sequentially coherent" as well as pairs that are incoherent, and ask the raters to rate the imageability of the sentences. The authors find that the imageability scores were "higher and faster when a pair of events was sequentially coherent rather than when the pair described two sequentially incoherent events".

Ramakrishna and Narayanan (2020) note that simple aggregations of word-level scores for psycholinguistic norms, such as imageability, may not best capture the sentence-level score, and define a sentence-level estimation which takes into account the relationships between the word-level norms. When evaluating the fusion model on predicting the imageability norm, the authors find that their estimations have smaller error in comparison to baselines, such as the average of word-level scores.

Kastner et al. (2021) create a system to generate image captions where the imageability and length of the caption is controllable. As a part of this, they develop an approach to estimate the imageability of a caption by weighting word-level scores by their position and relationships within the parse tree of the caption. This is used to create an imageability embedding to differentiate features of highly imageable captions and lowly imageable captions. Similarly to Ramakrishna and Narayanan (2020), the approach for defining an imageability score for a longer text involves a novel combination of existing word-level imageability scores.

DATASETS

For word-level imageability scores, which form the basis of our novel imageability score functions, we use the MRCPD+ lexicon, which contains imageability scores for 116,151 English words and phrases (Liu et al. 2014).

In order to determine pairs of words that are compatible within the same image and develop an imageability score function that considers pairwise

coherence, we use the ConceptualCaptions dataset from GoogleAI (Sharma et al. 2018). This dataset consists of pairs of images and corresponding captions that are filtered from web pages. The full training set for the image captioning task contains 3,318,333 examples and 51,201 unique tokens; in our work we use a validation set of 15,840 captions from the dataset. Since image captions describe an actual image, we note that words that co-occur in the same image caption can be used together to coherently describe the same image.

In order to help determine the impact of a word based on its frequency, we use the English Word Frequency dataset from Kaggle (Tatman, 2019). This dataset consists of words and the number of times they appear in the Google Web Trillion Word Corpus, for the 333,333 most frequent words.

APPROACH

In this section, we propose several novel ways to determine an imageability score for a text, utilizing combinations of word-level imageability scores. A first baseline method (**CI**) for determining the imageability score for a longer text, which utilizes only the word-level imageability scores from the MRCPD+, is the count of highly imageable words (words with an individual score above a threshold, for example, 0.7). In a simple sense, this is based on a hypothesis that if a text contains a higher number of imageable words, then the text evokes a stronger image.

A second baseline method (**HI**) for determining the imageability score for a longer text, which also utilizes only the word-level imageability scores from the MRCPD+, is the imageability score for the highest imageability individual word in the text. This is based on the idea that the image evoked by the text is related to the image that is evoked by the most imageable word.

We extend this second baseline method by considering additional words to just the most imageable word, and apply word embeddings, image caption data, and word frequency data. Initially, we design an imageability function (**DIS**) that considers the highest imageability word, as well as other highly imageable words at a discounted rate. Given the maximum word-level imageability score m, a threshold t, and a discount factor $d < 1$, the imageability score for the text is computed as follows: for the i^{th} word-level score s in the vector of all word-level scores:

if $s \geq m - t$, add $s * d^i$ to the score;

otherwise, if $m-t \geq s \geq m-2t$, subtract $s * d^i$ from the score.

Next, we apply clustering of word embedding vectors to identify sets of words in a text that may be related to the same image. We design an imageability function (**CLS**) to take into account consistency of words in a text and their relatedness to a main image, with the idea that imageability is higher if the main image is reinforced by other words in the text, e.g. adjectives, and lower if there are conflicting images. Given the maximum word-level imageability score m and a threshold t, we identify all words with imageability greater than or equal to m - t. Next, we obtain the FastText word vectors of these words, apply principal component analysis to reduce the dimensionality of the vectors to two, and cluster them using k-means clustering

(Bojanowski et al. 2017). We identify the cluster containing the highest imageability word (this is treated as the "main" image), and for all words within this cluster, we increase the score proportionally to the imageability s of the word and inversely proportionally to the distance d to the cluster centroid. This results in the updated imageability score of $score + s * (2-d)$. For words outside of the "main" cluster, we decrease the score proportionally to s and d, *i.e.*, $score-s * d$.

We note that the above function assumes that semantic relatedness of words within a text, as captured by word embeddings, is tied to whether these words are related in terms of imagery, which may not be the case. As a result, we design a second score (**CAP**) that considers relatedness of words in imagery, by utilizing image caption data. The general idea is that two words are compatible in an image if they can be found together in an image caption; and, if compatible words are found in the same text, then the imageability of the text should be greater. First, we create a word co-occurrence dictionary from image caption data. Using a validation set of 15,840 captions from the ConceptualCaptions dataset, we first remove all stop words and punctuation (Sharma et al. 2018). Next, we calculate the skip-bigram frequency for each pair of words, across all captions, to obtain co-occurrence counts for 10,620 unique words. Given two words, u and v, we define *compatibility(u, v)* as the ratio of co-occurrence of u with v and the total co-occurrence of u with all other words. The imageability score for a sentence is computed as a sum of compatibility-tuned imageability scores for a subset W of the most imageable words within this sentence, i.e., those with scores within a given threshold t from the maximum. A compatibility-tuned score for a word u with score s is obtained using the formula $s_c = s * \Pi(1 + 10 * compatibility(u, v))$, where the product is computed over all v in W.

Lastly, we develop an impact score for words and texts, aligning with our goal in relating imageability to influence. We exploit the notion that less commonly used words carry more information and thus may have a greater impact on an audience. Consequently, we combine imageability data from the MRCPD+ with word frequency data from the English Word Frequency dataset to define an impact score for individual words (Tatman, 2019). For any word w, the impact score $i(w)$ is defined as $s(w) * g(f(w))$ where s is the word-level imageability score, f is the frequency of w, and g is a monotonically decreasing function. We bound the range of g between -1 and 1, such that the most frequent word "the" has a score of -1 and words with lower frequency approach scores of 1; for example, $g(f(w)) = 1 + ln(1-\gamma f(w))$, where γ is a corpus dependent constant. Given this definition, we can create a word-level impact score dictionary, replacing the original imageability scores.

RESULTS

We compare the above imageability score functions and simple baseline functions in correlation between tweet imageability and number of retweets using a subset of tweets relating to the 2017 French presidential elections from Kaggle (Daignan, 2017). We extract 31,917 unique tweets from

April 9, 2017, to April 15, 2017, count the number of retweets for each tweet by considering duplicated tweets starting with "RT" (indicating a retweet), and filter the 738 tweets that have ten or more retweets. Of these tweets, the maximum number of retweets is 1,298, the mean number of retweets is 50.83, and the median number of retweets is 21. For these tweets, we compute our imageability score functions and investigate the relationship between the estimated imageability and the number of retweets.

Figure 1 shows graphs of number of retweets vs. imageability score function values for each of the imageability score functions and baselines described in the previous section.

Table 1 shows a statistically significant weak positive correlation between the imageability score computed using several functions discussed above and number of retweets. The functions that utilize the context between words have generally a slightly stronger relationship (except for the word embedding clustering-based function). Based on the trends in the figures, we note that tweets that receive considerably higher numbers of retweets (for example, greater than 400), will have between one and five highly imageable

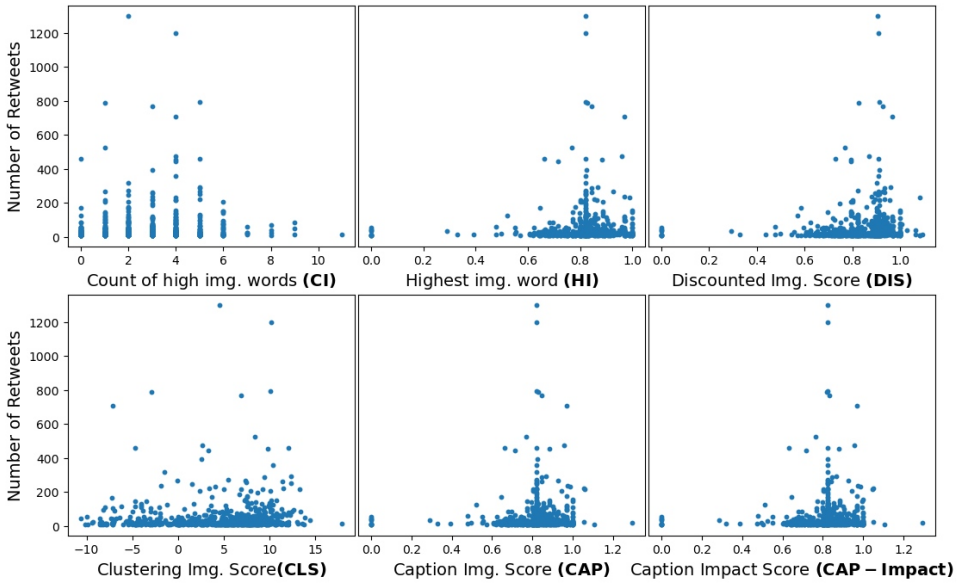

Figure 1: Number of retweets vs. imageability score functions for the six functions described in the previous section. For each plot, the x-axis is the imageability score and the y-axis is the number of retweets. **(CAP-Impact)** is the caption imageability score function with impact scores substituted for imageability scores.

Table 1. Spearman correlation coefficient and p-value for the relationship between the number of retweets vs. imageability score function values for each of the functions.

Imageability Score Function	(CI)	(HI)	(DIS)	(CLS)	(CAP)	(CAP-Impact)
Spearman	0.1067	0.1232	0.1439	0.0840	0.1331	0.1361
p-value	0.0037	0.0008	8.7169e-5	0.0224	0.0003	0.0002

words, with the highest word-level score of 0.6 or more. Intuitively, it means that using highly imageable words in a tweet does not guarantee it will receive many retweets; however, if a tweet receives many retweets, then it is more likely that it contains highly imageable language. We hypothesize, based on the trends, that the mean number of retweets for tweets within various continuous ranges of imageability scores increases for higher imageability score (**HI**) ranges. We test this by splitting the tweets into groups based on ranges of imageability scores, and computing the mean number of retweets for each group of tweets. Figure 2 shows the mean number of retweets vs. highest imageability score range when we consider six groups of tweets. We note that, on average, tweets with a higher imageability score receive more retweets.

Next, we consider case studies involving tweets related to specific topics or events, noting that content novelty and currency is a factor in the retweet rate. We consider three topics that are frequently discussed throughout the course of the 2017 French elections: "Le Pen", "Macron", and "France", and identify tweets relating to these topics by finding tweets with related hashtags: #LePen and #MarineLePen for "Le Pen", #Macron, #macron, and #EmmanuelMacron for "Macron", and #France and #france for "France". The count of unique tweets and overall mean number of retweets for each topic are shown in Table 2.

Figure 3 shows the mean number of retweets vs. highest imageability score (**HI**) range when we consider 10 groups of tweets. We note that across separate topics, tweets within higher ranges of imageability scores receive more retweets on average compared to tweets within lower ranges.

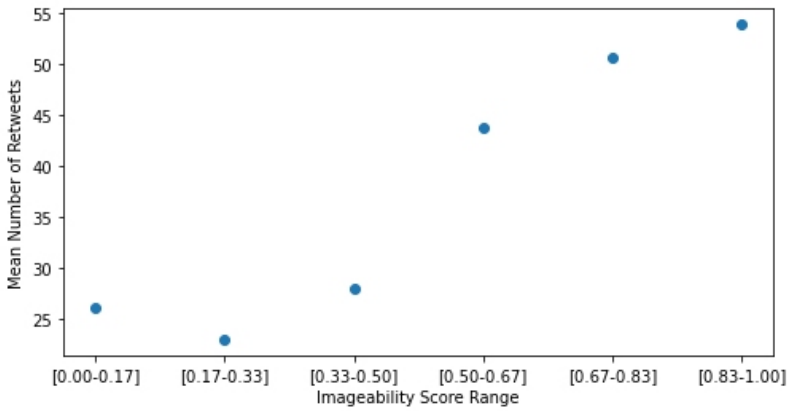

Figure 2: Mean number of retweets vs. highest imageability score range when we consider six groups of tweets.

Table 2. Count of unique tweets and overall mean number of retweets for the three topics "Le Pen", "Macron", and "France".

Topic	"Le Pen"	"Macron"	"France"
# Unique Tweets	1,293	1,541	1,082
Mean Number of Retweets	15.3387	10.9280	12.4436

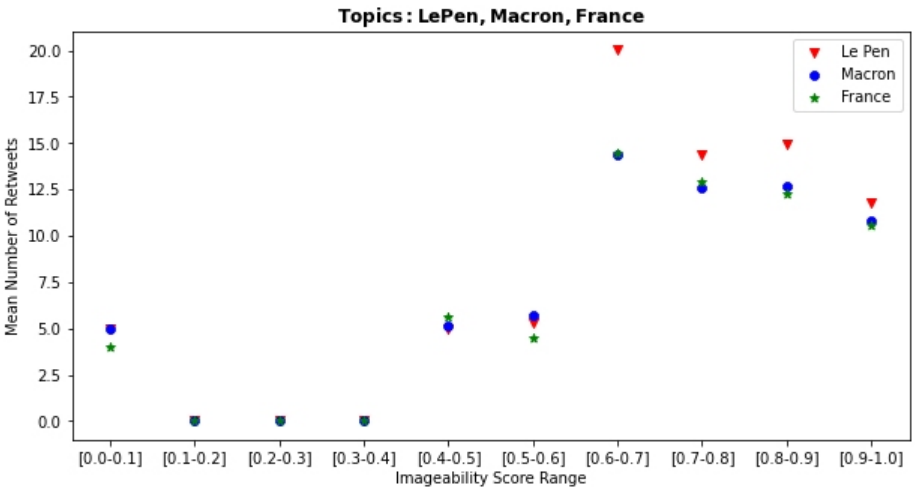

Figure 3: Mean number of retweets vs. highest imageability score range when we consider 10 groups of tweets, for the "Le Pen", "Macron", and "France" topics.

Lastly, we consider the "Macron email leaks" topic, which corresponds to a specific event that decays in popularity over time. We find 1,296 tweets related to the event by filtering tweets with the hashtags #MacronLeaks, #MacronGate, #macronleaks, and #macrongate within six days of the event, and split these tweets into four time frames. Table 3 shows the number of unique tweets and average retweet rate for tweets within each time frame.

Table 3. Number of unique tweets and average retweet rate for tweets within each time frame related to the "Macron email leaks" topic.

Time Frame	Days 1–2	Day 3	Days 4–5	Day 6
# Unique Tweets	1,024	393	123	18
Mean Number of Retweets	32.4229	26.7786	18.4309	5.8889

We note that over time, the event is tweeted about less frequently, and the retweet rate, on average, decays. Nonetheless, within each time frame, we find that the relationship between the imageability score range and the mean number of retweets still holds.

Ultimately, we find that the imageability score of messages is positively related to the mean number of retweets, when normalized for topic and novelty; thus, imageable language makes messages potentially more influential.

HUMAN ANNOTATION EXPERIMENT

With the correlation between tweet imageability and the number of retweets thus established, we wish to verify that our automated imageability functions align with human perception. To do so, we manually annotate a small

number of tweets for imageability to spot-check the quality of the imageability score function as an estimate of sentence-level imageability. We collected a subset of approximately 30,000 tweets relating to the 2022 French presidential elections, computed the imageability scores for each tweet using the function (**DIS**), and then split the tweets into three subsets: one containing tweets with scores [0, 0.7), one containing tweets with scores [0.7, 0.89), and one containing tweets with scores [0.89, Max]. From each subset, we randomly select 20 tweets, for a total of 60 tweets. The tweets from the lowest imageability subset are labeled as 1, the tweets from the mid-range imageability subset are labeled as 2, and the tweets from the highest imageability subset are labeled as 3. The tweets were then shuffled, with the labels hidden.

Four untrained annotators were given the following guidelines to annotate each tweet for imageability.

- For each message in your annotator sheet, answer the following:
  - *On the discrete scale of 1 (low imageability (including no image)) to 3 (high imageability), to what degree does the following message invoke a coherent image in your mind?*
  - *In a few words, describe the image(s) invoked, or write "None" if no specific image is invoked.*

After an initial tryout, the annotators were encouraged to compare their results and discuss their understanding of the first question. As a result, the following instruction was added: If the image forms immediately during or after reading the message use 3; if it takes a few seconds (∼5 sec) to form use 2; otherwise use 1.

For the four human raters, we computed Cohen's Kappa pairwise as a measure of inter-annotator agreement on the imageability annotation task. The Kappa scores, shown in Table 4, indicate low to moderate inter-annotator agreement and vary between pairs of annotators, reflecting the difficulty of the task. Additionally, we compute Cohen's Kappa between the scores from the imageability score function and each of the human annotators. We see that for some annotators (B and C), the system has a higher agreement in comparison to the other annotators, whereas, for some annotators (A and D), the system has the second highest agreement with the annotator.

Lastly, we consider the quality of the imageability score function for tweets for which human annotator agreement is high. From the 60 tweets, we filter

Table 4. Cohen's Kappa scores for each pair of human annotators and the system ratings. Scores highlighted in green indicate highest agreement with the annotator in the row.

	Ann. A	**Ann. B**	**Ann. C**	**Ann. D**	**System**
Ann. A		0.247659	0.159456	0.404389	0.3250
Ann. B	0.247659		0.157373	0.079365	0.3
Ann. C	0.159456	0.157373		0.134021	0.4
Ann. D	0.404389	0.079365	0.134021		0.15
System	0.3250	0.3	0.4	0.15	

the 33 tweets for which the majority (at least 3 out of 4) of the annotators agreed on the score. For these 33 messages, we find that the scores from the imageability score function agree with the majority for 23 messages (69.7%). As a result, we conclude that the imageability score function (DIS) is a reasonable estimate for the imageability of a tweet when the imageability is highly agreed upon by human raters.

CONCLUSION

In this paper, we introduced and tested several functions for computing imageability scores for short texts such as tweets. These functions utilize word embeddings, image caption data, and word frequency data to account for pairwise coherence between words and the impact of words on a target audience. Our experiments show that these imageability score functions are reasonable estimates for the overall imageability of a short text. Additionally, we find that tweets that are more imageable, on average, receive a higher number of retweets, and thus are potentially more influential. Further research is required to confirm this finding on a larger scale and across different topic domains.

ACKNOWLEDGMENT

This paper is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001121C0186. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA or the U.S. Government.

REFERENCES

Bakshy, E., Hofman, J., Mason, W., and Watts, D. (2011). Everyone's an Influencer: Quantifying Influence on Twitter. *4th ACM Int. Conf. on Web Search and Data Mining, WSDM 2011* (pp. 65–74).

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the ACL, 5*, p. 135–146.

Broadwell, G., Boz, U., Cases, I., Strzalkowski, T., Feldman, L., Taylor, S., Shaikh, S., Liu, T., Cho, K., and Webb, N. (2013). Using Imageability and Topic Chaining to Locate Metaphors in Linguistic Corpora. *6th international conference on Social Computing, Behavioral-Cultural Modeling and Prediction* (pp. 102–110).

Cha, M., Haddadi, H., Benevenuto, F. and Gummadi, K. (2010) "Measuring User Influence in Twitter: The Million Follower Fallacy", *International AAAI Conference on Web and Social Media*, 4(1), pp. 10-17. doi: 10.1609/icwsm.v4i1.14033.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *The Quarterly Journal of Experimental Psychology Section A, 33*(4), p. 497–505.

Daignan, J. (2017). French presidential election: Extract from twitter about the french election. https://www.kaggle.com/datasets/jeanmidev/french-presidential-election

Kastner, M., Umemura, K., Ide, I., Kawanishi, Y., Hirayama, T., Doman, K., Deguchi, D., Murase, H., and Satoh, S. (2021). Imageability- and Length-Controllable Image Captioning. *IEEE Access, 9*, p. 162951–162961.

Liu, T., Cho, K., Broadwell, G., Shaikh, S., Strzalkowski, T., Lien, J., Taylor, S., Feldman, L., Yamrom, B., Webb, N., Boz, U., Cases, I., and Lin, CS. (2014). Automatic Expansion of the MRC Psycholinguistic Database Imageability Ratings. *9th Int. Conf. on Language Resources and Evaluation* (pp. 2800–05). (ELRA).

Liu, T., Cho, K., Strzalkowski, T., Shaikh, S., and Mirzaei, M. (2016). The Validation of MRCPD Cross-language Expansions on Imageability Ratings. *10th Int. Conf. on Language Resources and Evaluation* (pp. 3748–3751) (ELRA).

Madden-Lombardi, C., Jouen, A.-L., Dominey, P. F., and Ventre-Dominey, J. (2015). Sequential coherence in sentence pairs enhances imagery during comprehension: An individual differences study. *PLoS ONE, 10*(9), Article e0138269. https://doi.org/10.1371/journal.pone.0138269

Paivio, A. (1969). Mental imagery in associative learning and memory. *Psychological Review, 76*, p. 241–263.

Paivio, A., Yuille, J. C., and Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology, 76*(1, Pt. 2), 1–25. https://doi.org/10.1037/h0025327

Ramakrishna, A, and Narayanan, S. (2020). Sentence level estimation of psycholinguistic norms using joint multidimensional annotations. *CoRR, abs/2005.10232*.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. (2018). Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. *56th Meeting of the Assoc. for Computational Linguistics* (pp. 2556–2565). ACL.

Tatman, R. (2019). English Word Frequency. Available at: https://www.kaggle.com/datasets/rtatman/english-word-frequency